LLM による日本語複単語表現の字義的用例生成

大橋美緒　木山朔　凌志棟　小町守
一橋大学
{mio, hajime, shito, komachi}@scl.sds.hit-u.ac.jp

概要

コーパスでは出現頻度が低い逐語的解釈をもつ日本語の複単語表現（Multiword Expressions: MWEs）について、大規模言語モデルが字義通りの用法の例文を生成できるかを検証する。JMWEL 辞書と JParaCrawl コーパスの両方に出現する MWE を対象とし、辞書に基づく語義記述や非字義的用例の提示有無を変えた複数の制御条件のもとで、GPT-5 にプロンプトを与え、字義的用例を生成させる。その結果、対照的な非字義的情報を与えることで字義的用法の生成が安定し、字義的情報のみを与えた場合や、手がかりを与えない場合と比べて、例文の品質が向上することが明らかになった。

1　はじめに

複単語表現（Multiword Expressions; MWEs）とは、同一の表層形でも字義的定義と、意味が構成要素から合成的に導けない非字義的定義を取り得る表現を指す [1]。MWE は文脈によって字義的な意味か非字義的な意味かの判断が必要となり、自然言語処理においても重要な課題である [2]。このような両義性を扱う上での大きな要因として、コーパスにおいて非字義的用法に大きく偏ったコーパス分布がある。

本研究では、この問題が特に顕著に現れる日本語の MWE に焦点を当てる。コーパス構築のためには一定量の字義的用例が必要であるが、日本語は、英語と比べて、自由に再利用可能なテキストデータが大幅に少ない [3]。そのうえ、日本語では語順の自由度が高く、字義通り用例が表層的に捉えにくい一方、慣用的用例は固定的に現れやすいため、利用可能なデータ量はさらに少ない [4]。

MWE 理解に関するモデル評価は人手注釈付き用例に依存するため、用例が字義通りもしくは非字義通りのいずれかに偏っていると MWE 理解の適切な評価や資源の構築が困難になる [5]。多くの日本語 MWE において字義的用例は非字義的用例よりもはるかに稀であり、観測される比率は主として用例が収集されたコーパスにおける使用頻度を反映している。実際、日本語 MWE について、字義的用例と非字義的用例を体系的に対照する資源は存在しない。

言語資源	用例	字義的 vs. 非字義的	カバレッジ
OpenMWE	✓	✓	△
JMWEL	✗	✗	✓
提案手法	✓	✓	✓

表 1　既存の日本語 MWE リソースとの比較：(i) 用例の有無、(ii) 字義的用法と非字義的用法の対比、(iii) カバレッジ

このギャップを踏まえ、本研究では、大規模言語モデル（LLM）を用い、日本語 MWE の字義的用例を制御されたプロンプト条件下で生成する。図 1 は、字義的用例生成および評価のための本研究の枠組みを示している。本研究では、JMWEL と JParaCrawl の両方に出現する MWE を対象とし、自動評価および人手評価を用いて生成例を評価する [6]。表 1 に、本研究と先行研究との違いについてまとめる。

本研究の主な貢献は以下の通りである。

(i) 制御された LLM プロンプトにより、字義的用例を生成する手法を提案する。
(ii) 日本語 MWE をケーススタディとして、既存の語彙資源およびコーパスに含まれる MWE に本手法を適用し、生成された字義的用例の性質を分析する。
(iii) LLM-as-a-judge および小規模な人手評価を用いて生成例を評価し、多くの例が解釈可能で字義通りであることを示す。
(iv) 日本語 MWE の字義的用例と非字義的用例を対照する言語資源を構築し、公開を予定している[1)]。

2　関連研究

LLM に基づく用例生成　近年、LLM は、辞書情

1）https://github.com/SDS-NLP/JMWE-parallel

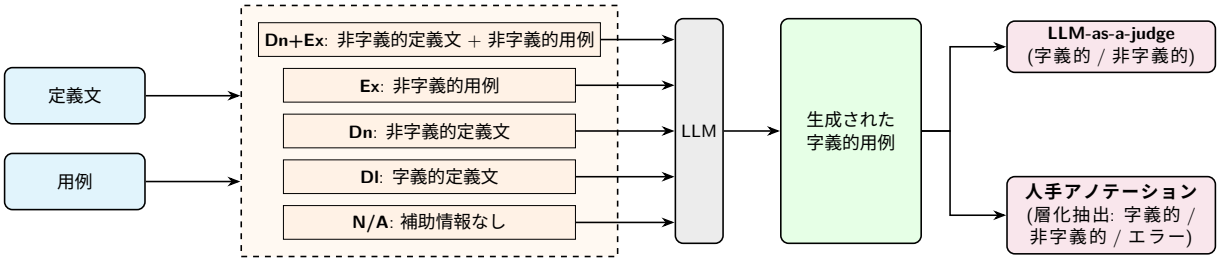

図 1　日本語複単語表現の字義的用例生成に関する本研究のフレームワーク。定義文および非字義的用例の利用可否が異なる 5 条件に基づく制御付きプロンプトと、自動評価および人手評価から構成される。

報に条件づけられた用例を生成するために用いられている。Cassotti と Tahmasebi [7] は、辞書の定義文と単語の書かれた時期に基づいて生成された用例が高い品質を達成し、英語における通時的な意味変化研究のための語義別データとして利用可能であることを示した。しかし、この研究は、非構成的で構成要素間の強い相互作用を示すことの多い MWE には、直接的には適用されてこなかった。本研究は、LLM に基づく用例生成を日本語 MWE へと拡張し、既存コーパスでは非字義的解釈が支配的な表現に対して、字義的用例の生成を求めた場合に、モデルがどのように振る舞うかを分析する。

MWE 資源　OpenMWE [8] は、コーパスから抽出した日本語 MWE に対して、literal（字義的）や idiomatic（非字義的）といった用法ラベルを付与した資源であるが、収録されている表現や用例の数には限りがあり、また元コーパスでの出現状況に強く依存するため、用例の分布に偏りが生じやすい。JMWEL [9] は、形態素分割や語形成情報、統語パターンなどを含む詳細な言語学的アノテーションを備えている一方で、字義的用法と非字義的用法を対にした用例を含まず、文脈情報もほとんど付与されていないため、文脈に基づく NLP タスクへの利用には制約がある。総じて、既存の MWE 資源は、字義的用法と非字義的用法の体系的かつ対照的な分析を十分に支援しておらず、そのことが代替的アプローチの必要性の示唆している。

3　実験手法

本研究では、日本語の MWE についてその字義的用例を LLM が生成するタスクを定義する。対象となる MWE が与えられると、LLM は異なる種類の入力情報を条件として、その表現の字義的用例を生成する。本研究のフレームワークは、用例生成と評価の段階から構成される。図 1 に、本研究のフレームワークを示す。本研究の目的は、特定のモデル性能の評価ではなく、字義的用例生成においてどのような情報が有効であるかを明らかにすることである。

	定義文 （非字義的）	定義文 （字義的）	用例 （非字義的）
Dn+Ex	✓	–	✓
Ex	–	–	✓
Dn	✓	–	–
Dl	–	✓	–
N/A	–	–	–

表 2　各プロンプト条件において提供される情報を示す。“定義文” は意味記述（字義的または非字義的）を指し、“用例” は非字義的用例を指す。

生成段階では、対象 MWE に加え、用法記述や MWE 関連アノテーションなどの外部資源に基づく語彙情報を異なる組み合わせでプロンプト条件として LLM に与える。本研究では、入力情報の組み合わせが異なる 5 つの条件を定義する。生成タスクは固定したまま入力情報のみを体系的に変化させ、条件間で比較することが本フレームワークの中核を成す。表 2 に、各プロンプト条件を示す。以下では、「手を引く」を例としたときの各条件の内容を示す。

- **Dn+Ex**　MWE の表現形式に加え、非字義的定義文（例：「比喩的に、関係から撤退すること、関与を断つこと」）と、非字義的用例（例：「市場から手を引く」）を提示する。
- **Ex**　MWE の表現形式と、非字義的用例（例：「市場から手を引く」）のみを提示し、意味定義は与えない。
- **Dn**　MWE の表現形式と、非字義的定義文（例：「比喩的に、関係から撤退すること、関与を断つこと」）のみを提示し、用例は与えない。
- **Dl**　MWE の表現形式と、字義的定義文（例：「文字通り、人の手を引くこと、先導すること」）を提示する。
- **N/A**　MWE の表現形式のみを提示し、用例および定義文はいずれも与えない。

	MWE	生成要求数	生成数
Dn+Ex	338	1,690	1,283
Ex	338	1,690	1,323
Dn	449	2,245	1,855
Dl	358	1,790	1,624
N/A	449	2,245	1,389

表 3　各プロンプト条件における対象 MWE 数および生成文数の統計を示している。

評価段階では、生成に用いなかった LLM による自動評価と人手評価を行う。このように生成と評価を分類することで、生成モデル由来のバイアスを抑えつつ、字義的用例の品質を評価する。

4　実験設定

本研究では、OpenAI API により提供される GPT-5 [2] を生成モデルとして用い、各対象 MWE について、各条件につき字義的用例を 5 文ずつ、合計 25 文生成させた．これらの対象 MWE に対して、定義文情報として『日本国語大辞典 第二版』を用い、449 件の対象 MWE について辞書項目に基づき、用例を削除した上で、字義的定義文と非字義的定義文を自動的に分類し、プロンプト条件に応じてこれらの意味記述を GPT-5 への入力に含めるか否かを切り替えた。また、用例については、JMWEL をインデックスとして用い、JParaCrawl から対象 MWE を含む文を抽出した。

表 3 は、各プロンプト条件における対象 MWE の数および生成された用例の数を示している。なお、5 つのプロンプト条件ごとに MWE の数が異なるのは字義的用例、非字義的定義文、字義的定義文が入力となるが、MWE ごとに入力となる用例や定義文が存在せず条件を満たさないものがあるためである。

生成された字義的用例の品質を、自動評価と人手評価の両方を用いて評価する。自動評価として LLM-as-a-judge アプローチを採用し、Gemini-2.5-pro[3] を用いて各生成文を字義的定義文または非字義的定義文に分類し、字義的と判断された文の割合（literal ratio）を算出した。各条件の字義的用例の比率は、判定モデルによって字義的と判断された文の割合として算出し、本研究では補助的指標として用いた。一方、人手評価では「エラー」ラベルを導入し、各文を「字義的」「非字義的」「エラー」のいずれかに分類した。人手評価は層化された 250 文（各プロンプト条件につき 50 文）のサンプルに対して実施した。人手評価は、日本語母語話者 2 名、すなわち第一著者と大学院生によって実施された。

抽出された各文は、字義的用例または非字義的用例のいずれかとして手動でアノテーションしたが、詳細なガイドラインに従い、独立に行った。アノテータ間一致度は Cohen's κ により測定し、人手評価における literal ratio は、両アノテータの判断が一致した文のみに基づいて算出した。

2）　gpt-5-2025-08-07
3）　gemini-2.5-pro、2025 年 11 月時点

		literal ratio		
	生成率	自動	人手	κ
Dn+Ex	74%	0.98	0.88	0.29
Ex	78%	0.99	0.88	0.12
Dn	76%	0.99	0.94	0.19
Dl	83%	0.53	0.36	0.72
N/A	62%	0.99	0.82	0.03
Overall	**75%**	**0.90**	**0.78**	**0.41**

表 4　各条件における生成率、自動・人手評価による字義的用例と判断された文の比率、およびアノテータ間一致度（Cohen's κ）を示す。

5　実験結果

本研究では、生成成功率、字義性判断（自動および人手）という 2 つの側面から生成結果を報告する。

5.1　生成成功率

まず、モデルが要求された形式で出力を正しく生成できているかを検証する。表 4 に各条件の成功率を示し、空文字列、不正な JSON、または必要な文数を満たさない出力を失敗として扱った。N/A は最も低い成功率を示しており、完全に制約のないプロンプトが不安定な生成挙動を引き起こすことを裏付けている。一方、定義文や非字義的用例を与える条件（Dn+Ex、 Ex、Dn、Dl）は、概してより高い成功率を示した。この結果は、プロンプトに何らかの言語的ヒントを含めることが、意味の制御に加えて生成過程の安定化にも有効であることを示唆している。

5.2　字義性と一致率

自動評価　表 4 は、Gemini-2.5-pro による自動評価の結果を示しており、条件 Dn+Ex、 Ex、Dn、N/A は約 0.98–0.99 と同程度に高い literal ratio を示し、最

小限の意味的文脈下では生成文が一貫して字義的と判断された。一方、Dl は成功率が高いにもかかわらず、literal ratio が大幅に低く、字義的用例の生成の不安定性が示された。

人手評価　表 4 には、各条件における人手評価に基づく literal ratio と Cohen's κ を示している。人手評価に基づく literal ratio は条件間で大きく異なり、Dn+Ex、Ex、Dn、N/A では 0.82–0.94 と高い値を示した一方、Dl では 0.36 と著しく低い。これは、LLM の内部知識における字義的意味記述が、プロンプトで与えた辞書の定義文によって阻害される傾向にあるためと考えられる。また、アノテータ間一致度は条件によって異なり、Dl では高い一致度（$\kappa = 0.72$）が得られた一方、N/A では著しく低い一致度（$\kappa = 0.03$）となった。全体としての一致度は中程度であった（$\kappa = 0.41$）。

6　議論

量的分析に加え、どのような MWE の生成が失敗するかを確認するために人手評価の詳細な質的分析を実施した。特に、慣用的意味が強く慣習化した MWE では字義的解釈が実質的に困難であった。

人手評価におけるアノテータ間一致を見ると、完全一致ケースのうち 145 件が 字義的、33 件が非字義的、3 件がエラーと判断された。

非字義的と判定された用例を (1)–(3) に示す。

(1) 彼は会議で秘密を口を滑らせた。
(2) 手術の説明を聞き、私は腹を括った。
(3) 取材では歯に衣を着せないコメントが目立った。

(1) の「口を滑らせる」 は身体部位が行為者から独立して振る舞うという前提が不自然さを生み、(2) の「腹を括る」 は理論上可能でも現代日本語では極めて不自然であり、(3) の「歯に衣を着せぬ」 は字義的解釈が概念的にほぼ不可能であるなど、字義的解釈可能性には大きな差がある。しかし、(1)–(3) の文は、人手評価で一貫して非字義的と判断された。その要因は、自然な字義的文脈を構築できず、とくに字義的解釈可能性の低い MWE において、強く慣習化された慣用的用例に寄せたためと推測される。

この点は、生成モデル固有の問題というよりも、MWE における字義的解釈可能性の性質そのものに起因すると考えられる。というのも、MWE の字義的解釈の可能性は多くの場合、連続的な性質を持ち、事象の概念化、動詞と名詞の共起選好、語彙意味の拡張といった微妙な制約に左右される。定性的分析が示すように、理論上は字義通りの定義が存在する場合であっても、その使用が自然であるとは限らず、強い制約を受けることがある。これは、辞書では字義的用法と非字義的用法が明確に区別されているものの、実際の言語使用においてはどちらとも解釈できる事例があることが背景にある。

また、エラーと判定された用例は、文法的には正しいが、MWE ごとの語彙的・意味的制約に違反している。用例を (4)–(6) に示す。

(4) 鬼ごっこで相手の足を引っ張る。
(5) 筋トレの下降局面で力をゆるめつつ、
　ゆっくり息を抜く。
(6) 紙コップの口を揃えて重ねる。

(4) の「足を引っ張る」では、文法的には正しくても、ゲームの進行に関する常識的な事象構造と整合しないため不自然さが生じる。(5) の「息を抜く」では、行為自体は解釈可能でも、共起頻度と慣用性が低いため、かなり不自然であるが、慣習的なコロケーションに注意を向けるよう促すことで、他の語彙的制約よりもプロンプトによる部分的緩和が期待できる。(6) の「口を揃える」では無生物への拡張自体は可能であるものの、紙コップという対象が典型的形状・機能の期待に反するため、不自然であると言える。この点で、語彙意味拡張に関する制約は、プロンプトへの反映が難しい。

以上より、字義的用例生成がすべての MWE において一様に達成可能とはいえず、むしろ言語的制約を考慮したプロンプト戦略およびタスク設計が必要性だといえる。

7　おわりに

本稿では、コーパスからの抽出が困難な日本語 MWE の字義的用例を補完するため、生成ベースの枠組みを従来のコーパス資源と統合することで、より高品質な MWE データセットの構築を支援し、LLM による字義的意味と非字義的意味の対照の表現や活用の実態を明らかにすることを可能にした。

今後の展望として、より大規模な人手アノテーション付きデータを用いた判定の較正や、異なる評価モデルでの傾向の分析が挙げられる。

謝辞

本研究成果は、国立研究開発法人情報通信研究機構 (NICT) の委託研究 “自動翻訳の精度向上のための「マルチモーダル情報の外部制御可能なモデリング」の研究開発” によって得られたものである。

参考文献

[1] Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. Discriminative Lexical Semantic Segmentation with Gaps: Running the MWE Gamut. **Transactions of the Association for Computational Linguistics**, Vol. 2, pp. 193–206, 2014.

[2] Timothy Baldwin and Su Nam Kim. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, **Handbook of Natural Language Processing**, pp. 267–292. Chapman and Hall/CRC, 2010.

[3] Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. The VNC-Tokens Dataset. 2008.

[4] Chikara Hashimoto and Daisuke Kawahara. Construction of an Idiom Corpus and its Application to Idiom Identification based on WSD Incorporating Idiom-Specific Features. In Mirella Lapata and Hwee Tou Ng, editors, **Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing**, pp. 992–1001, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[5] Chikara Hashimoto, Satoshi Sato, and Takehito Utsuro. Detecting Japanese Idioms with a Linguistically Rich Dictionary. **Language Resources and Evaluation**, Vol. 40, No. 3–4, pp. 243–252, 2006.

[6] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus. In **Proceedings of The 12th Language Resources and Evaluation Conference**, pp. 3603–3609, Marseille, France, May 2020. European Language Resources Association.

[7] Pierluigi Cassotti and Nina Tahmasebi. Sense-specific Historical Word Usage Generation. **Transactions of the Association for Computational Linguistics**, Vol. 13, pp. 690–708, 2025.

[8] Chikara Hashimoto and Daisuke Kawahara. Compilation of an Idiom Example Database for Supervised Idiom Identification. **Language Resources and Evaluation**, Vol. 43, No. 4, pp. 355–384, 2009.

[9] Kosho Shudo, Toshifumi Tanabe, and Masahito Takahashi. Overview and Current Status of the Japanese Multiword Expression Lexicon JMWEL: Focusing on verbal multiword expressions. In **Proceedings of the 2018 Workshop on Language Resource Utilization**, pp. 601–610. National Institute for Japanese Language and Linguistics, 2018.

A プロンプト

A.1 SYSTEM prompt.

あなたは日本語文作成支援ツールです。
- 各 MWE（慣用句）について、『文字通りの用法の日本語文』を作成するかどうかを判断し、必要なら文を生成します。
- 文字通りの用法が存在すると判断した場合は、その MWE を文字通りの意味で使った自然な日本語文を、最大 {N_LIT} 文まで作成してください。
- 文字通りの用法が存在しない、または辞書上ほぼ比喩的・抽象的な意味に限られていると判断した場合でも、無理に文を作らなくて構いません。

注：以下のブロックのみ条件ごとに異なる
Dn+Ex: - 入力として与えられる非文字通り用例（example_nonlit）と辞書定義（dict_def_nonlit）から MWE の意味を理解し、
Ex: - 入力として与えられる非文字通り用例（example_nonlit）のみから MWE の意味を推測し、
Dn: - 入力として与えられる辞書定義（dict_mean_non_literal）から MWE の比喩的・慣用的な意味を理解し、
Dl: - 入力として与えられる『文字通りの意味の辞書定義（dict_mean_literal）』に基づき、
N/A:- 入力として与えられる情報は MWE の表層形だけです。辞書定義や非文字通り用例など、その他のヒントとなる情報は一切利用できないものと考えてください。

文字通りの用法があるかどうかを慎重に判断し、has_literal に true/false を設定してください。
- 非慣用・非比喩。「ように／かのように」を含めてはいけない。
- 各文は句点「。」で終え、句点は 1 つのみとする。
- 同じ MWE について生成する複数の文は、できるだけ意味や構文が重複しないようにする。
- 出力は説明を含めず JSON のみ。

A.2 USER prompt header.

以下のデータに基づき、各 MWE について文字通りの用法の日本語文を生成してください。
- 各 MWE について、文字通りの用法が存在すると判断した場合は literal_list に最大 {N_LIT} 文まで返してください。
- 文字通りの用法が存在しないと判断した場合は、has_literal を false、literal_list を空配列 [] にして構いません。
出力形式:

```
{
  "items": [
    {"mwe":   "...",  "has_literal":   true,
"literal_list": ["...", "..."] },
    {"mwe":   "...",  "has_literal":   false,
"literal_list": []},
    ...
  ]
}
```

データ: